(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7599554号
(P7599554)

(45)発行日　令和6年12月13日(2024.12.13)

(24)登録日　令和6年12月5日(2024.12.5)

(51)国際特許分類

Ｇ０６Ｔ　13/40　(2011.01)		Ｇ０６Ｔ	13/40
Ｇ１０Ｌ　25/63　(2013.01)		Ｇ１０Ｌ	25/63

ＦＩ

請求項の数　16　（全13頁）

(21)出願番号	特願2023-514901(P2023-514901)
(86)(22)出願日	令和3年9月2日(2021.9.2)
(65)公表番号	特表2023-540535(P2023-540535A)
(43)公表日	令和5年9月25日(2023.9.25)
(86)国際出願番号	PCT/US2021/048870
(87)国際公開番号	WO2022/051497
(87)国際公開日	令和4年3月10日(2022.3.10)
審査請求日	令和5年4月27日(2023.4.27)
(31)優先権主張番号	63/074,310
(32)優先日	令和2年9月3日(2020.9.3)
(33)優先権主張国・地域又は機関	米国(US)
(31)優先権主張番号	17/396,664
(32)優先日	令和3年8月7日(2021.8.7)

最終頁に続く

(73)特許権者	310021766 株式会社ソニー・インタラクティブエンタテインメント 東京都港区港南１丁目７番１号
(74)代理人	100105924 弁理士　森下 賢樹
(72)発明者	カウシィク、ラクシュミシュ アメリカ合衆国、カリフォルニア州 ９４４０４、サン マテオ、ブリッジポイント パークウェイ ２２０７
(72)発明者	クマール、サケット アメリカ合衆国、カリフォルニア州 ９４４０４、サン マテオ、ブリッジポイント パークウェイ ２２０７
審査官	中田 剛史

最終頁に続く

(54)【発明の名称】　テキストと音声を用いたフェイシャルアクションユニットの自動生成によるフェイシャルアニメーション制御

(57)【特許請求の範囲】

【請求項１】

装置であって、

少なくとも１つのプロセッサであって、

コンピュータ化されたアバターの顔の画像を識別すること、

前記アバターに関連する第１のモダリティデータであって、音声を含む前記第１のモダリティデータを識別すること、

前記アバターに関連する第２のモダリティデータであって、テキストを含む前記第２のモダリティデータを識別すること、

前記アバターの前記顔の前記画像をアニメーション化するのに役立つ、前記第１のモダリティデータに少なくとも部分的に基づく第１の情報及び前記第２のモダリティデータに少なくとも部分的に基づく第２の情報を受け取ること、及び

前記第１の情報及び前記第２の情報の両方に従って前記アバターの前記顔をアニメーション化すること、

のための命令で構成された、前記少なくとも１つのプロセッサ、

を含み、

前記命令は少なくとも１つの機械学習（ＭＬ）モデルにアクセスして前記アバターの前記顔をアニメーション化することを実行可能であり、前記ＭＬモデルは前記アバターの前記顔の画像から導出されたアンカーポイントを受け取ってフェイシャルアクションユニット（ＦＡＵ）を生成し、前記音声及び前記テキストを前記ＦＡＵに関連付けることを実行

可能であり、
前記命令は、前記音声から抽出された音素境界を少なくとも部分的に使用して前記音声及び前記テキストを時間的に互いに整合させることを実行可能であり、前記命令は、前記音声から抽出された前記音素境界を使用して感情及び感情の程度を前記音声及び前記テキストと時間において整合させることをさらに実行可能である、前記装置。
【請求項２】
前記命令は、前記第１及び第２のモダリティデータから前記感情を導出することを実行可能である、請求項１に記載の装置。
【請求項３】
前記情報は、前記感情から生成された時間で整合された単語レベルの感情確率に少なくとも部分的に基づく、請求項２に記載の装置。
【請求項４】
方法であって、
第１のテキスト及び第１の音声の両方に従って単語を話すアニメーション化される第１の顔の画像を生成すること、
前記第１のテキスト及び前記第１の音声を時間において整合させること、及び
前記第１のテキスト及び前記第１の音声に従って、第１の単語を話す前記第１の顔の前記画像をアニメーション化すること、
を含み、
前記第１の顔の前記画像は、前記第１の音声における単語のスライディングウィンドウを処理することによって少なくとも部分的にアニメーション化され、前記第１の音声における単語１～Ｎは第１の感情に関連付けられ、前記第１の顔の前記画像は前記第１の感情に従ってアニメーション化され、前記第１の音声における単語２～Ｎ＋１は第２の感情に関連付けられ、前記第１の顔の前記画像は前記第２の感情に従ってアニメーション化され、Ｎは２よりも大きい整数である、前記方法。
【請求項５】
既知の単語を話すアニメーション化された顔のトレーニングセットを使用して、機械学習（ＭＬ）モデルをトレーニングすること、
前記第１のテキスト及び前記第１の音声を前記ＭＬモデルに入力すること、
前記ＭＬモデルの出力に従って、前記第１の顔の前記画像をアニメーション化すること、
前記第１のテキストから感情と情緒を検出すること、
前記第１のテキストを前記第１のテキストを表す音声と整合させて、整合されたテキスト／音声をレンダリングすること、及び
前記感情、前記情緒、及び前記整合されたテキスト／音声を前記ＭＬモデルに入力すること、
を含む、請求項４に記載の方法。
【請求項６】
ターゲット感情を前記ＭＬモデルに入力すること
を含む、請求項５に記載の方法。
【請求項７】
前記ＭＬモデルからフェイシャルアクションを表す第１の確率を受け取ること
を含む、請求項５に記載の方法。
【請求項８】
前記ＭＬモデルから感情を表す第２の確率を受け取ること
を含む、請求項７に記載の方法。
【請求項９】
前記第１及び第２の確率の一方または両方を使用して、フェイシャルアクションユニット（ＦＡＵ）を確立すること
を含む、請求項８に記載の方法。
【請求項１０】

前記ＦＡＵに従って前記第１の顔の前記画像をアニメーション化すること
を含む、請求項９に記載の方法
【請求項１１】
アセンブリであって、
アニメーション化されたコンピュータアバターを提示するように構成された少なくとも１つのディスプレイ、
機械学習（ＭＬ）モデルを実行するための命令で構成された少なくとも１つのプロセッサであって、前記命令は、
前記アバターが話す音声を示すテキストを受け取ること、
音声を受け取ること、
前記音声から抽出された音素境界を少なくとも部分的に使用して前記テキスト及び前記音声を時間において整合させること、
前記ＭＬモデルを使用して前記テキスト及び前記音声を処理し、フェイシャルアクションユニット（ＦＡＵ）を生成すること、及び
前記ＦＡＵに従って前記コンピュータアバターをアニメーション化すること、
に対して実行可能である、前記少なくとも１つのプロセッサ、
を含む、前記アセンブリ。
【請求項１２】
前記命令は、
前記テキストから感情及び情緒を検出すること、
前記テキストを、前記テキストを表す音声に整合させて、整合されたテキスト／音声をレンダリングすること、及び
前記感情、前記情緒、及び前記整合されたテキスト／音声を前記ＭＬモデルに入力すること、に対して実行可能である、請求項１１に記載のアセンブリ。
【請求項１３】
前記命令は、
ターゲット感情を前記ＭＬモデルに入力することに対して実行可能である、請求項１２に記載のアセンブリ。
【請求項１４】
前記命令は、
前記ＭＬモデルからフェイシャルアクションを表す第１の確率を受け取ること
に対して実行可能である、請求項１１に記載のアセンブリ。
【請求項１５】
前記命令は、
前記ＭＬモデルから感情を表す第２の確率を受け取ること
に対して実行可能である、請求項１４に記載のアセンブリ。
【請求項１６】
前記命令は、
前記第１及び第２の確率を使用して前記ＦＡＵを確立することに対して実行可能である、請求項１５に記載のアセンブリ。
【発明の詳細な説明】
【技術分野】
【０００１】
本願は、概して、コンピュータシミュレーション及び他のアプリケーションでのテキスト及び音声を使用したフェイシャルアクションユニット（ＦＡＵ）の自動生成によるフェイシャルアニメーション制御に関する。
【背景技術】
【０００２】
コンピュータゲームのアバターなどのコンピュータ化された画像の顔は、シミュレーションのプレイ中にアニメーション化され、リアルな効果が得られる。本明細書で理解され

るように、アーティストは通常、人間の顔のそれぞれの筋肉点の収縮または弛緩をエミュレートすることによって、フレームごとに顔のそれぞれの部分をアニメーション化するために使用されるフェイシャルアニメーションユニット（ＦＡＵ）を苦労して作成する必要がある。

【発明の概要】

【０００３】

装置は、少なくとも１つのプロセッサであって、コンピュータ化されたアバターの顔の画像を識別すること、という命令で構成された、少なくとも１つのプロセッサを含む。命令は、機械学習（ＭＬ）エンジンへのアバターに関連する第１のモダリティデータを入力すること、アバターに関連する第２モダリティデータをＭＬエンジンに入力すること、及びアバターの顔の画像をアニメーション化するのに役立つＭＬエンジンから出力を受け取ることに対して実行可能である。命令はまた、出力に従ってアバターの顔をアニメーション化することに対して実行可能である。

【０００４】

いくつかの実施形態で、出力は、複数のフェイシャルアクションユニット（ＦＡＵ）を含み、各ＦＡＵは、顔の画像のそれぞれの部分に関係する。非限定的な例で、第１のモダリティデータがテキストを含み、第２のモダリティデータは音声を含む。

【０００５】

非限定的な例では、命令は、ＭＬエンジンを使用して、第１及び第２のモダリティデータから感情アクション情報を導出することに対して実行可能である。出力は、感情アクション情報から生成された時間で整合された単語レベルの感情確率に少なくとも部分的に基づき得る。

【０００６】

別の態様では、方法が、既知の単語を話すアニメーション化された顔のトレーニングセットを使用して、機械学習（ＭＬ）モデルをトレーニングすることを含む。方法はまた、第１のテキストに従って単語を話すアニメーション化される第１の顔の画像を生成することを含む。方法は、第１のテキストをＭＬモデルに入力すること、及びＭＬモデルの出力に従って、第１のテキストによって示される第１の単語を話す第１の顔の画像をアニメーション化することをさらに含む。

【０００７】

例示的な実施形態では、方法は第１のテキストから感情と情緒を検出すること、第１のテキストを第１のテキストを表す音声と整合させて、整合されたテキスト／音声をレンダリングすること、及び感情、情緒、及び整合されたテキスト／音声をＭＬモデルに入力することを含み得る。方法はまた、ターゲット感情をＭＬモデルに入力することを含み得る。方法の例示的実装形態では、フェイシャルアクションを表すＭＬモダリティから第１の確率を受け取ること、感情を表すＭＬモデルから第２の確率を受け取ること、及び第１及び第２の確率を使用して、フェイシャルアクションユニット（ＦＡＵ）を確立することを含み得る。次いで方法は、ＦＡＵに従って第１の顔の画像をアニメーション化することを含むことができる。

【０００８】

別の態様では、アセンブリが、アニメーション化されたコンピュータアバターを提示するように構成された少なくとも１つのディスプレイ、及び機械学習（ＭＬ）モデルを実行するための命令で構成された少なくとも１つのプロセッサを含む。命令は、アバターが話す音声を示すテキストを受け取ること、ＭＬモデルを使用してテキストを処理し、フェイシャルアクションユニット（ＦＡＵ）を生成すること、及びＦＡＵに従ってコンピュータアバターをアニメーション化することに対して実行可能である。

【０００９】

本願の詳細は、その構造と動作との両方について、添付の図面を参照すると最もよく理解でき、図面において、類似の参照符号は、類似の部分を指す。

【図面の簡単な説明】

【００１０】
【図１】本原理による例を含む例示的なシステムのブロック図である。
【図２】例示的なコンピュータ化されたアバターの顔のスクリーンショットである。
【図３】例示的なフローチャート形式で例示的な全体的なロジックを示す。
【図４】機械学習（ＭＬ）エンジンのトレーニングプロセスを示す。
【図５】ＭＬエンジンの推論プロセスを示す。
【発明を実施するための形態】
【００１１】
本開示は、概して、限定されることなく、コンピュータゲームネットワークなどの家電（ＣＥ）デバイスネットワークの態様を含むコンピュータエコシステムに関する。本明細書のシステムは、クライアントコンポーネントとサーバコンポーネントとの間でデータが交換され得るように、ネットワークを通じて接続され得るサーバコンポーネント及びクライアントコンポーネントを含み得る。クライアントコンポーネントは、Ｓｏｎｙ　ＰｌａｙＳｔａｔｉｏｎ（登録商標）などのゲームコンソールまたはＭｉｃｒｏｓｏｆｔ（登録商標）もしくはＮｉｎｔｅｎｄｏ（登録商標）もしくは他の製造者によって作成されたゲームコンソール、仮想現実（ＶＲ）ヘッドセット、拡張現実（ＡＲ）ヘッドセット、ポータブルテレビ（例えば、スマートテレビ、インターネット対応テレビ）、ラップトップ及びタブレットコンピュータなどのポータブルコンピュータ、ならびにスマートフォン及び以下で議論される追加の実施例を含む他のモバイルデバイスを含む、１つ以上のコンピューティングデバイスを含み得る。これらのクライアントデバイスは、様々な動作環境で動作し得る。例えば、クライアントコンピュータのいくつかは、実施例として、Ｌｉｎｕｘ（登録商標）オペレーティングシステム、Ｍｉｃｒｏｓｏｆｔ（登録商標）のオペレーティングシステム、またはＵｎｉｘ（登録商標）オペレーティングシステム、またはＡｐｐｌｅ，Ｉｎｃ．（登録商標）もしくはＧｏｏｇｌｅ（登録商標）によって制作されたオペレーティングシステムを採用し得る。これらの動作環境は、Ｍｉｃｒｏｓｏｆｔ（登録商標）もしくはＧｏｏｇｌｅ（登録商標）もしくはＭｏｚｉｌｌａ（登録商標）によって作成されたブラウザ、または以下で議論されるインターネットサーバによってホストされるウェブサイトにアクセスできる他のブラウザプログラムなど、１つ以上の閲覧プログラムを実行するために使用され得る。また、本原理による動作環境を使用して、１つ以上のコンピュータゲームプログラムを実行し得る。
【００１２】
サーバ及び／またはゲートウェイは、インターネットなどのネットワークを通じてデータを受信及び送信するようにサーバを構成する命令を実行する１つ以上のプロセッサを含み得る。あるいは、クライアント及びサーバは、ローカルイントラネットまたは仮想プライベートネットワークを通じて接続することができる。サーバまたはコントローラは、Ｓｏｎｙ　ＰｌａｙＳｔａｔｉｏｎ（登録商標）などのゲームコンソール、パーソナルコンピュータなどによってインスタンス化され得る。
【００１３】
クライアントとサーバとの間でネットワークを通じて情報を交換し得る。この目的及びセキュリティのために、サーバ及び／またはクライアントは、ファイアウォール、ロードバランサ、テンポラリストレージ、及びプロキシ、ならびに信頼性及びセキュリティのための他のネットワークインフラストラクチャを含むことができる。１つ以上のサーバは、ネットワークメンバーにオンラインソーシャルウェブサイトなどの安全なコミュニティを提供する方法を実装する装置を形成し得る。
【００１４】
プロセッサは、アドレスライン、データライン及び制御ラインなどの様々なライン、並びにレジスタ及びシフトレジスタによって論理を実行することができる、シングルチッププロセッサまたはマルチチッププロセッサであってよい。
【００１５】
一実施形態に含まれるコンポーネントは、他の実施形態では、任意の適切な組み合わせ

で使用することができる。例えば、本明細書に記載される、及び／または図で示される様々なコンポーネントのいずれもは、組み合わされ、交換され、または他の実施形態から除外されてもよい。

【００１６】

「Ａ、Ｂ及びＣのうちの少なくとも１つを有するシステム」（同様に「Ａ、ＢまたはＣのうちの少なくとも１つを有するシステム」及び「Ａ、Ｂ、Ｃのうちの少なくとも１つを有するシステム」）は、Ａ単独、Ｂ単独、Ｃ単独、Ａ及びＢを一緒に、Ａ及びＣを一緒に、Ｂ及びＣを一緒に、ならびに／またはＡ、Ｂ及びＣを一緒に有するシステムなどを含む。

【００１７】

ここで、具体的に図１を参照すると、本原理による、上述され、以下でさらに説明される例示的なデバイスのうちの１つ以上を含み得る例示的なシステム１０が示されている。システム１０に含まれる例示的なデバイスのうちの第１のデバイスは、限定されることなく、テレビチューナ（同等に、テレビを制御するセットトップボックス）を備えたインターネット対応テレビなどのオーディオビデオデバイス（ＡＶＤ）１２などの家電（ＣＥ）デバイスである。代替として、ＡＶＤ１２は、また、コンピュータ制御型インターネット対応（「スマート」）電話、タブレットコンピュータ、ノートブックコンピュータ、ＨＭＤ、ウェアラブルコンピュータ制御デバイス、コンピュータ制御型インターネット対応ミュージックプレイヤ、コンピュータ制御型インターネット対応ヘッドフォン、インプラント可能な皮膚用デバイスなどのコンピュータ制御型インターネット対応インプラント可能デバイス、などであってもよい。それにも関わらず、ＡＶＤ１２は、本原理を実施する（例えば、本原理を実施するように他のＣＥデバイスと通信し、本明細書に記載される論理を実行し、本明細書に記載されるいずれかの他の機能及び／または動作を行う）ように構成されることを理解されたい。

【００１８】

したがって、このような原理を実施するために、ＡＶＤ１２は、図１に示されているコンポーネントの一部または全てによって確立することができる。例えば、ＡＶＤ１２は、１つ以上のディスプレイ１４を備えることができ、このディスプレイは、高解像度もしくは超高解像度「４Ｋ」またはそれ以上の解像度のフラットスクリーンによって実装されてもよく、ディスプレイのタッチを介したユーザ入力信号を受信するためにタッチ対応であってもよい。ＡＶＤ１２は、本原理に従ってオーディオを出力するための１つ以上のスピーカ１６、及び可聴コマンドをＡＶＤ１２に入力してＡＶＤ１２を制御するためのオーディオ受信機／マイクロホンなどの、少なくとも１つの追加入力デバイス１８を含み得る。例示的なＡＶＤ１２は、また、１つ以上のプロセッサ２４の制御の下、インターネット、ＷＡＮ、ＬＡＮなどの少なくとも１つのネットワーク２２を通じて通信するための１つ以上のネットワークインタフェース２０を含み得る。また、グラフィックプロセッサ２４Ａが含まれていてもよい。したがって、インタフェース２０は、限定されることなく、Ｗｉ－Ｆｉ（登録商標）送受信機であり得て、このＷｉ－Ｆｉ（登録商標）送受信機は、限定されることなく、メッシュネットワーク送受信機などの無線コンピュータネットワークインタフェースの実施例である。プロセッサ２４は、その上に画像を提示するようにディスプレイ１４を制御すること及びそこから入力を受信することなど、本明細書に記載されるＡＶＤ１２の他の要素を含むＡＶＤ１２が本原理を実施するように、制御することを理解されたい。さらに、ネットワークインタフェース２０は、有線もしくは無線のモデムもしくはルータ、または、例えば、無線テレフォニ送受信機もしくは上述したＷｉ－Ｆｉ（登録商標）送受信機などの他の適切なインタフェースであってよいことに留意されたい。

【００１９】

上記のものに加えて、ＡＶＤ１２はまた、例えば、別のＣＥデバイスに物理的に接続する高解像度マルチメディアインタフェース（ＨＤＭＩ（登録商標））ポートもしくはＵＳＢポート、及び／またはヘッドフォンを通してＡＶＤ１２からユーザにオーディオを提供するためにＡＶＤ１２にヘッドフォンを接続するヘッドフォンポートなどの１つ以上の入力ポート２６を含んでもよい。例えば、入力ポート２６は、オーディオビデオコンテンツ

のケーブルまたは衛星ソース２６ａに有線でまたは無線で接続されてもよい。したがって、ソース２６ａは、別個のもしくは統合されたセットトップボックス、または衛星受信機であってよい。あるいは、ソース２６ａは、コンテンツを含むゲームコンソールまたはディスクプレイヤであってもよい。ソース２６ａは、ゲームコンソールとして実装されるとき、ＣＥデバイス４４に関連して以下で説明されるコンポーネントの一部または全てを含んでよい。

【００２０】

ＡＶＤ１２は、さらに、一時的信号ではない、ディスクベースストレージまたはソリッドステートストレージなどの１つ以上のコンピュータメモリ２８を含んでもよく、これらのストレージは、場合によっては、スタンドアロンデバイスとしてＡＶＤのシャーシ内で、またはＡＶプログラムを再生するためにＡＶＤのシャーシの内部もしくは外部のいずれかでパーソナルビデオ録画デバイス（ＰＶＲ）もしくはビデオディスクプレイヤとして、または取り外し可能メモリ媒体として具現化されてもよい。また、ある実施形態では、ＡＶＤ１２は、限定されることなく、携帯電話受信機、ＧＰＳ受信機、及び／または高度計３０などの位置または場所の受信機を含むことができ、位置または場所の受信機は、衛星もしくは携帯電話基地局から地理的位置情報を受信し、その情報をプロセッサ２４に供給し、及び／またはＡＶＤ１２がプロセッサ２４と併せて配置されている高度を決定するように構成される。コンポーネント３０はまた、通常、加速度計、ジャイロスコープ、及び磁力計の組み合わせを含み、ＡＶＤ１２の位置及び方向を３次元で決定する慣性測定ユニット（ＩＭＵ）によって実装されてもよい。

【００２１】

ＡＶＤ１２の説明を続けると、いくつかの実施形態では、ＡＶＤ１２は、１つ以上のカメラ３２を含んでよく、１つ以上のカメラは、サーマルイメージングカメラ、ウェブカメラなどのデジタルカメラ、及び／またはＡＶＤ１２に統合され、本原理に従って写真／画像及び／またはビデオを収集するようプロセッサ２４によって制御可能なカメラであってよい。また、ＡＶＤ１２に含まれるのは、Ｂｌｕｅｔｏｏｔｈ（登録商標）及び／または近距離無線通信（ＮＦＣ）技術を各々使用して、他のデバイスと通信するためのＢｌｕｅｔｏｏｔｈ（登録商標）送受信機３４及び他のＮＦＣ要素３６であってよい。例示的なＮＦＣ素子は、無線周波数識別（ＲＦＩＤ）素子であってもよい。

【００２２】

さらにまた、ＡＶＤ１２は、プロセッサ２４に入力を供給する１つ以上の補助センサ３７（例えば、加速度計、ジャイロスコープ、サイクロメータなどの運動センサ、または磁気センサ、赤外線（ＩＲ）センサ、光学センサ、速度センサ及び／またはケイデンスセンサ、ジェスチャセンサ（例えば、ジェスチャコマンドを検知するための）など）を含み得る。ＡＶＤ１２は、プロセッサ２４への入力をもたらすＯＴＡ（無線）ＴＶ放送を受信するための無線ＴＶ放送ポート３８を含み得る。上記に加えて、ＡＶＤ１２はまた、赤外線データアソシエーション（ＩＲＤＡ）デバイスなどの赤外線（ＩＲ）送信機及び／またはＩＲ受信機及び／またはＩＲ送受信機４２を含み得ることに留意されたい。電池（図示せず）は、電池を充電するために及び／またはＡＶＤ１２に電力を供給するために運動エネルギーを電力に変えることができる運動エネルギーハーベスタのように、ＡＶＤ１２に電力を供給するために提供され得る。

【００２３】

さらに図１を参照すると、ＡＶＤ１２に加えて、システム１０は、１つ以上の他のＣＥデバイスタイプを含み得る。一実施例では、第１のＣＥデバイス４４は、ＡＶＤ１２に直接送信されるコマンドを介して及び／または後述のサーバを通して、コンピュータゲームの音声及びビデオをＡＶＤ１２に送信するために使用することができるコンピュータゲームコンソールであり得る一方で、第２のＣＥデバイス４６は第１のＣＥデバイス４４と同様のコンポーネントを含み得る。図示の実施例では、第２のＣＥデバイス４６は、プレイヤによって操作されるコンピュータゲームのコントローラとして、またはプレイヤ４７によって装着されるヘッドマウントディスプレイ（ＨＭＤ）として構成され得る。図示の実

施例では、２つのＣＥデバイス４４、４６のみが示されているが、より少ないまたはより多くのデバイスが使用されてよいことは理解されよう。本明細書のデバイスは、ＡＶＤ１２について示されているコンポーネントの一部またはすべてを実装し得る。次の図に示されているコンポーネントのいずれかに、ＡＶＤ１２の場合に示されているコンポーネントの一部またはすべてが組み込まれることがある。

【００２４】

ここで、上述の少なくとも１つのサーバ５０を参照すると、サーバは、少なくとも１つのサーバプロセッサ５２と、ディスクベースストレージまたはソリッドステートストレージなどの少なくとも１つの有形コンピュータ可読記憶媒体５４と、サーバプロセッサ５２の制御下で、ネットワーク２２を通じて図１の他のデバイスとの通信を可能にし、実際に、本原理に従ってサーバとクライアントデバイスとの間の通信を容易にし得る少なくとも１つのネットワークインタフェース５６とを含む。ネットワークインタフェース５６は、例えば、有線もしくは無線モデムもしくはルータ、Ｗｉ－Ｆｉ送受信機、または、例えば、無線テレフォニ送受信機などの他の適切なインタフェースであってよいことに留意されたい。

【００２５】

したがって、いくつかの実施形態では、サーバ５０は、インターネットサーバまたはサーバ「ファーム」全体であってよく、「クラウド」機能を含んでもよく、「クラウド」機能を実行してもよく、システム１０のデバイスが、例えば、ネットワークゲームアプリケーションの例示的な実施形態においてサーバ５０を介して「クラウド」環境にアクセスし得るようにする。あるいは、サーバ５０は、図１に示されている他のデバイスと同じ部屋にある、またはその近くにある、１つ以上のゲームコンソール、または他のコンピュータによって実装されてもよい。

【００２６】

図２は、コンピュータアバターの顔２０２が提示される本明細書の任意のディスプレイなどのディスプレイ２００を示す。顔２０２は、人間の顔の対応する部分の弛緩または収縮をエミュレートするために、それぞれのフェイシャルアニメーションユニット（ＦＡＵ）によって要求されるようにアニメーション化され得る、多くのエミュレートされたアンカーポイント２０４（明確にするために４つだけが示されている）を有する。

【００２７】

図３は、ＦＡＵを自動的に生成して顔２０２をアニメーション化するために使用できる全体的なロジックを示す。本明細書で説明するロジックは、本明細書で説明するプロセッサのうちの任意の１つまたは複数によって実行することができる。

【００２８】

ブロック３００で始まり、機械学習（ＭＬ）エンジンが、図４を参照して説明されるように訓練される。ブロック３０２に進むと、顔のビデオ画像２０２が、例えばコンピュータシミュレーション、例えばコンピュータゲームにより生成され、顔に関連する対応するマルチモダリティデータがブロック３０４でＭＬエンジンに入力される。図５を参照してさらに十分に説明するように、マルチモダリティデータから、ＭＬエンジンは、ブロック３０６でアバターの顔２０２のエミュレートされた筋肉をアニメーション化するために使用される予測ＦＡＵを生成する。

【００２９】

図４は、ＭＬエンジンのトレーニングプロセスの例を示している。データベース４００からのデータのトレーニングセットがアクセスされる。データのトレーニングセットには、話しているときに感情を表現する顔の画像が含まれている。換言すれば、データベースは、正しい表現をアニメーション化しながら既知のテキストを話すアバターのグラウンドトゥルースアニメーションを含む。

【００３０】

入力されるトレーニング画像ごとに、画像のそれぞれのアンカーポイント２０４がブロック４０２で画像から抽出される。言い換えれば、ブロック４０２は、フェイシャルアク

ションユニット（ＦＡＵ）に使用される入力画像のアンカーポイントを生成する。また、音声信号及び音声信号が関連付けられているテキストは、ブロック４０４で抽出され、ブロック４０４での音声及びテキストがブロック４０２で抽出されたＦＡＵに関連付けられていることが理解される。

【００３１】

ブロック４０６は、以下にさらに開示するように、音声とテキストが互いに時間的に整合していることを示す。図４のグラフ４０７によって表されるように、音素境界は、対応するテキスト４０７Ｂの音声信号４０７Ａの音素境界から抽出される。このようにすると、ｘ軸に沿った各インクリメンタルブロックが音声によって生成された音素を表し、音素がインクリメンタルブロックの隣接する境界４０７Ｃ、４０７Ｄの間で定められる。

【００３２】

強制整合ブロック４０６は、音素境界を使用して、感情注釈ブロック４０８で感情を時間において整合し、エキスパートが感情で入力に注釈を付ける感情注釈ブロック４１０で感情を整合し、音声特徴ブロック４１２で音声の特徴を整合させる。したがって、対応する感情、情緒、及び音声の特徴は、抽出された音素境界を使用して整合される。以下でさらに説明するように、「情緒」は感情の程度、例えば肯定的または否定的であることを指すことができ、「感情」は話者の感情的な状態、例えば幸福、悲しみ、怒りなどを指すことができる。音声における単語のスライディングウィンドウは、この方法で処理できる、例えば、単語１～３に第１の感情／情緒で注釈を付けることができ、単語２～４に第２の感情／情緒で注釈を付けることができる、等々であることに留意されたい。音声特徴４１２は、入力の組み合わされた特性を表すために生成される。ウィグナー・ヴィル分布関数及び時間周波数フィルタリング技法を含む技法を使用して音声特徴を生成することができる。

【００３３】

ブロック４０８、４１０、及び４１２の生成物は、ＭＬ感情アクション生成モデル４１４に供給され、モデルをトレーニングする。また、モデル４１４への入力は、ブロック４０２からのＦＡＵと、モデル４１４をトレーニングするために使用される入力される（注釈付き）感情のグラウンドトゥルースであるターゲット感情４１６である。このようにすると、モデル４１４はデータベース４００からの複数のデータセットでトレーニングされ、ＦＡＵを対応する音声及びテキスト及びターゲット感情と相関させて、フェイシャルアクションアンカー確率４１８及び時間で整合させた単語レベル感情確率４２０を出力し、その後、ＦＡＵに従ってアンカーポイント２０４を適切にアニメーション化することによって、顔（例えば、図２の顔２０２）をアニメーション化する際に使用することができる。

【００３４】

図５に移る前に、感情は、カテゴリ及び次元的な感情の分類の一方または両方を使用してグラウンドトゥルース４１６と比較するために、入力トレーニングセットで検出され得る。カテゴリ分類は、幸福、悲しみ、怒り、恐怖などを含む感情カテゴリを生成する。次元的な分類は、感情価と覚醒度の観点から感情の次元を生成する。

【００３５】

図５は、モデル４１４によるその後の推論処理を示し、図５では５１０とラベル付けされている。アニメ化されたアバターが話すテキスト５００は、テキストを音声信号に変換するためにテキスト読み上げブロック５０２に入力される。さらに、または代わりに、テキストに対応する記録された、または生の人間の音声がブロック５０２に供給され得る。図４を参照して上で教示されたように整合されたものに対応するブロックの要素を時間的に整合させるために、図４のグラフ４０７と同じ特徴を有するグラフ５０７によって示されるように、図４のトレーニングプロセスについて説明されたのと同じ方法で、ブロック５０４で音声の特徴が抽出され、ブロック５０６でテキスト及び対応する音声から音素境界が抽出される。

【００３６】

ブロック５００からのアバターが話すテキストは感情及び情緒検出ブロック５０８に送

られ、図４からのトレーニングに従ってテキストから情緒及び感情を抽出する。ブロック５０８でテキストから抽出された感情及び情緒は、ブロック５０４からの音声の特徴及びブロック５０６からの整合されたテキスト／音声信号とともに、モデル５１０に入力される。所望であれば、ブロック５１２に示されるように、ターゲット感情がモデルに入力され得る。ターゲット感情は、ユーザが入力するか、機械学習を使用して入力されたテキストから導出される、注釈付きテキストの一部であってもよい。

【００３７】

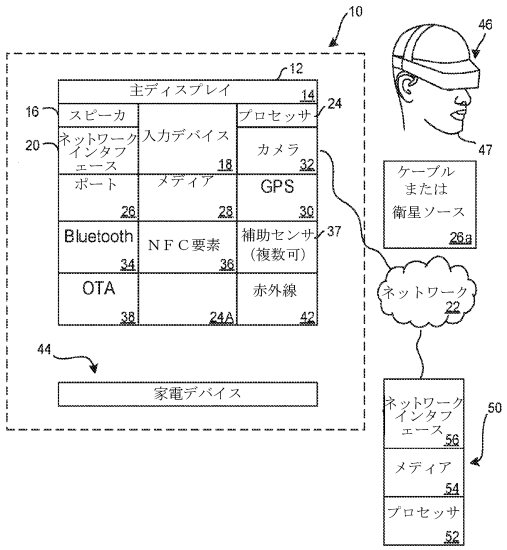
そのトレーニングに基づいて、モデル５１０は、受け取った入力から、各アンカーポイント２０４のフェイシャルアクションの確率５１２と、時間で整合されたテキスト／音声信号から導出された感情の確率５１６を出力する。これらの確率の一方または両方に基づいて、投影されたアンカー位置５１８（すなわち、ＦＡＵ）が得られ、各アニメーション化されたフレームが描写する顔の表情を確立する。ブロック５００で、ＦＡＵがアバターのアニメーションに適用され、アバターがテキスト入力を話すときに正しい表情を与える。

【００３８】

いくつかの例示的な実施形態を参照して本原理を説明したが、これらは限定することを意図しておらず、各種の代替的な構成が本明細書で特許請求される主題を実施するために使用されてよいことは理解されよう。

【図面】

【図１】



【図２】

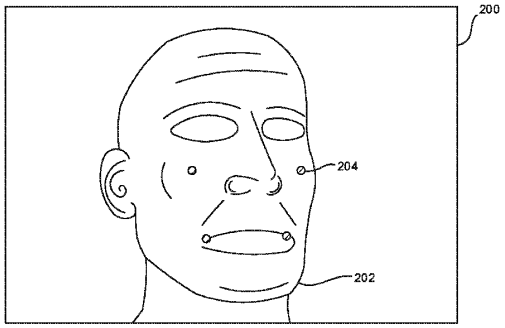
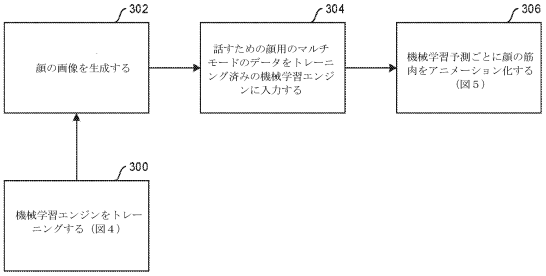
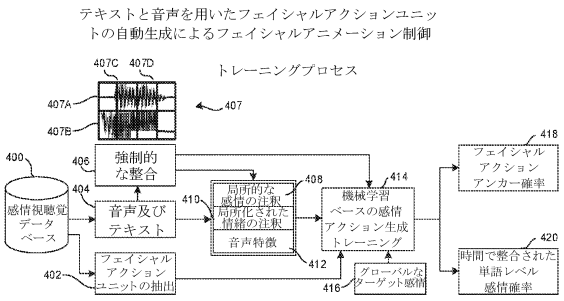


FIG. 2

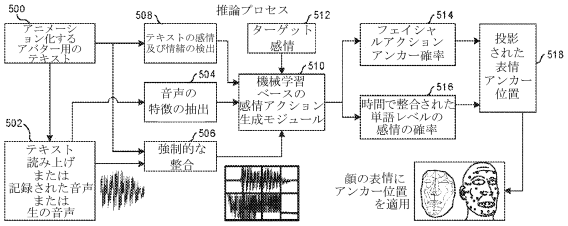
【図３】



【図４】



【図５】


500
アニメーション化するアバター用のテキスト
502
テキスト読み上げまたは記録された音声または生の音声
508
テキストの感情及び情緒の検出
504
音声の特徴の抽出
506
強制的な整合
推論プロセス
512
ターゲット感情
510
機械学習ベースの感情アクション生成モジュール
514
フェイシャルアクションアンカー確率
516
時間で整合された単語レベルの感情の確率
518
投影された表情アンカー位置
顔の表情にアンカー位置を適用

フロントページの続き

(33)優先権主張国・地域又は機関
米国(US)
(56)参考文献 国際公開第２０１９／１６０１００（WO，Ａ１）
特開２００７－０５８８４６（ＪＰ，Ａ）
特開２００５－３５２８９２（ＪＰ，Ａ）
(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｔ １３／４０
Ｇ１０Ｌ ２５／６３